Formation à l'analyse de données RNA-seq Galaxy

Exercices

Objectifs:

Cette formation a pour but de vous aider à traiter les séquences issues des SGS (Seconde Generation Sequencing) en particulier les platesformes Illumina (GAIIx, HiSeq). Vous y découvrirez les nouveaux formats de séquences, les biais connus et mettrez en œuvre des logiciels d'alignement épissé sur génome de référence, la recherche de nouveaux gènes, de nouveaux transcrits et la quantification de l'expression de ces gènes et transcrits.

Pour réaliser l'ensemble de ces exercices, connectez-vous avec votre utilisateur sur une plateforme Galaxy depuis un navigateur :

- http://sigenae-workbench.toulouse.inra.fr/
- http://migale.jouy.inra.fr/galaxy/

Les données que nous utiliserons sont accessibles à cette adresse : http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/data

Vous pouvez également utiliser un des comptes formation : anemone aster bleuet iris muguet narcisse pensee rose tulipe violette

Dans ce TP, les lignes en Italique correspondent à des outils à utiliser hors de Galaxy.

Exercice n°1: Data Quality

Quelques liens:
- EMBL-ENA (European Nucleotide Archive) : http://www.ebi.ac.uk/ena/

Étude des données à utiliser :
http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/data

Il s'agit des deux jeux de données reduits du chromosome 6 de la tomate.

- WT : Wild type, paired
- MT : Wild type, paired

Dans Galaxy :

- Créez un historique rnaseq
- Importez les données dans votre historique avec l'outil Upload File
- Changez le nom de chaque dataset (utiliser un nom court)
- Visualisez le contenu de chacun des fichiers pour vérifier leur chargement.

Analyse de la qualité des données:

- Utilisez FastQC en renommant le fichier de sortie à partir du nom du dataset
- Analysez les résultats

- Quelle est la longueur des lectures, quelle est la qualité du séquençage, regardez les résultats concernant les biais décrits lors du cours, lesquels retrouve-t-on ?

Nettoyage de données

- Utilisation de sickle avec les paramètres :
  - Minimal length of 20
  - Minimal mean quality of 20
  - No N in seq
  - No 5' trimming

- Renommez le fichier de sortie à partir du nom du dataset.

Vérifiez à nouveau la qualité des données nettoyées, quel jeu de données préférez-vous utiliser ?

<u>Exercice n°2</u>: alignement/visualisation

Quelques liens:

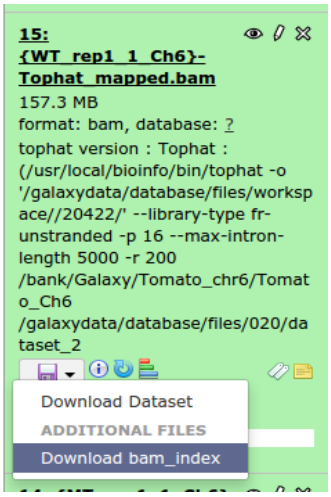
- Tophat: http://tophat.cbcb.umd.edu/
- Samtools: http://samtools.sourceforge.net/
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- FTP download de Ensembl: http://www.ensembl.org/info/data/ftp/index.html

Aujourd'hui nous allons nous focaliser sur l'alignement sans transcriptome de référence avec les paramètres de base. Pour lancer l'alignement, il vous faut une référence, vérifier si cette référence existe en utilisant tophat.

Si votre génome d'intérêt n'existe pas, veuillez faire une demande auprès du support.

- Lancez tophat en paired-end avec une taille d'insert de 200bp et une taille maximale d'introns de 5000bp pour les deux jeux de données nettoyés contre la référence nommée « Tomato Chr 6 »

- Utilisez « samtools flagstat » sur les fichiers nommés accepted_hits.bam pour un résumé des statistiques d'alignement.
- Téléchargez sur votre ordinateur les fichiers de résultats de topĥat (bam et bed) et le fichier d'indexation (bai)
- Renommez ces fichiers bam et bai avec les nom des runs



Visualisation des résultats :

- *Utilisez IGV pour visualiser les résultats sur votre poste de travail.*
- *Lancez IGV depuis « download » du site web de la formation (en bas de la page):*
  http://www.broadinstitute.org/software/igv/download

- *Chargezles annotations (fichier gtf mis a disposition dans*
  http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/data/reference/ITAG_pre2.3_gene_models_Ch6.gtf
- *Chargez les .bam, .bed*

- *Explorez l'interface, utiliser le clic-droit (pour visualiser toutes les isoformes, les paires de lectures associées....)*

- *Regardez les régions montrées dans le cours ainsi que les régions suivantes :*

  SL2.40ch06:22,595-27,402
  SL2.40ch06:38,480,842-38,484,938
  SL2.40ch06:10,694,176-10,704,838
  SL2.40ch06:9,839,693-9,862,815
  Solyc06g009140.2.1

Pour info :
Lors d'une analyse future, il se pourrait que les noms des chromosomes soient différents entre ceux intégrés dans IGV et ceux de votre gtf
Dans ce cas, il faut utiliser un fichier de correspondance; par exemple : Zv9_alias.tab à mettre dans le répertoire :
/home/...../igv/genomes
http://www.broadinstitute.org/software/igv/LoadData/#aliasfile
Ce fichier devra alors contenir les correspondances entre les noms utilisés par IGV et les noms de votre GTF :

1 chr1
2 chr2
3 chr3
4 chr4
5 chr5...
….

Exercice 3 : Recherche de nouveaux transcrits :

- Créez un fichier bam contenant tous les alignements de tous les échantillons (avec samtools merge)

- Lancez cufflinks avec le GTF de référence, avec le fichier bam complet (afin d'obtenir un gtf complet correspondant à nos échantillons).

Durant le traitement, familiarisez vous avec le format de fichier GTF grâce aux questions suivantes :

- Visualisez le contenu du fichier GTF dans galaxy.
- Utilisez l'utilitaire (Ensembl GTF statistics)
  - Combien y a-t-il de gènes ?
  - Combien y a-t-il de transcrits ?

Quand cufflinks est terminé :

- *Chargez la nouvelle annotation (Assembled transcript) dans IGV. (Téléchargez le fichier et l'ouvrir avec IGV)*
- *Explorez cette nouvelle annotation de transcrits.*

- Pour vous guider dans l'exploration, comparez, dans galaxy, la nouvelle annotation avec celle de référence (cuffcompare).
- Regardez les statistiques générales.
- Extrayez du fichier refmap, les lignes dont la troisième colonne n'est pas '=' et allez *voir pour chaque type de transfrag un exemple dans IGV, puis retournez voir les zones citées dans l'exercice 2.*

Exercice n°4: mesure d'expression brute au niveau gènes/transcrits :

Quantification au niveau gènes à l'aide du gtf de référence et Htseq-count :

- Lancez htseq count sur chaque dataset sachant que les données ne sont pas strand-spécifique et que l'on veut les intersections de gène non vide.

- Changez le type en **tabular** (et non txt)

- Fusionnez les résultats pour obtenir une matrice de comptage à l'aide de Merge tabulated files on first column

Quantification au niveau transcrits à l'aide du gtf de référence et Feature count :

- Lancer feature count avec le GTF de référence sur l'ensemble des échantillons (utiliser les paramètres définis dans le tutoriel)
- Testez également le comptage au niveau des gènes avec cet utilitaire.

Quelques outils pour l'analyse d'expression différentielle :

Il existe toute une batterie de package Bioconductor disponible pour l'analyse différentielle de données RNA-seq. Ces outils sont en plein développement et ne sont pas encore mature (difficulté dans le choix de la méthode à appliquer, choix de normalisation, évolution rapide des versions avec souvent de forts changements d'une version à l'autre...)

1. DESEq:
   http://bioconductor.org/packages/release/bioc/html/DESeq.html
2. EdgeR
   http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html